

A Graph-Based Approach to Analyze Flux-Balanced Pathways in Metabolic Networks

Mona Arabzadeh¹, Morteza Saheb Zamani¹, Mehdi Sedighi¹,
and Sayed-Amir Marashi²

¹Department of Computer Engineering and Information Technology,
Amirkabir University of Technology, Tehran, Iran and

²Department of Biotechnology, College of Science, University of Tehran, Tehran, Iran.

{m.arabzadeh,szamani,msedighi}@aut.ac.ir, marashi@ut.ac.ir

Abstract

An Elementary Flux Mode (EFM) is a pathway with minimum set of reactions that are functional in steady-state constrained space. Due to the high computational complexity of calculating EFMs, different approaches have been proposed to find these flux-balanced pathways. In this paper, an approach to find a subset of EFMs is proposed based on a graph data model. The given metabolic network is mapped to the graph model and decisions for reaction inclusion can be made based on metabolites and their associated reactions. This notion makes the approach more convenient to categorize the output pathways. Implications of the proposed method on metabolic networks are discussed.

Keywords: Elementary Flux Mode (EFM); Graph Data Model; Metabolic Network

1 Introduction

Genome-scale metabolic network models are among the well-studied models in biotechnology. The reconstruction of these networks is possible by collecting the gene-protein-reaction information from related genomic data and literature [1]. It is important to explore biologically relevant pathways in genome-scale metabolic networks. Forcing constraints to a reconstructed biochemical network results in the definition of achievable cellular functions [2]. Mathematical representation of constraints are as *flux-balance* constraints (e.g., conservation of flux) which means the network should be at the steady-state condition, and *flux bounds* which limit numerical ranges of network parameters and coefficients such as the minimum and maximum range of fluxes for each reaction. Flux Balance Analysis (FBA), a method to predict the optimal growth rate of a certain species when grown on particular set of metabolites [3], and Elementary Flux Mode (EFM) analysis, an approach to decompose a network to minimal functional pathways [4], are among the approaches used in constraint-based analysis of metabolic networks.

EFMs [5, 6] have been used in several biological applications such as bio-engineering [7], phenotypic characterization [8], drug target prediction [9] and strain design [10]. The incorporation of kinetic analysis into EFMs enables a

more complete description of cellular functions for which kinetics play a dominant role [11].

Several methods were introduced for finding EFMs based on double-description method [5]. Double-description is a technique to enumerate all extreme rays of a polyhedral cone. An improved approach to the primary method was introduced in which the null-space of the stoichiometric matrix is used instead of the matrix itself to generate EFM candidates [12, 13, 14]. Various effective computational approaches have been proposed to speed-up previous methods for computing EFMs [15, 16]. Some of these approaches led to the development of computational tools such as *Metatool* [17] and *EFMtool* [16]. Besides, methods based on linear programming have been proposed that explore a set of EFMs with specific properties, such as *K*-shortest EFMs [18], or EFMs with a given set of target reactions [19]. A new set of methods based on graph-theory, [20, 21] tries to overcome the scalability problem of the double-description-based techniques.

There is an additional challenge of extracting biological properties from large set of EFMs [11]. Even for networks with the same number of reactions and metabolites, knowing the number of EFMs in one network cannot necessarily help in finding the number of EFMs in the other one, as they have different connectivities and topologies. This fact emphasizes the inherent structural information reflected by EFMs [22] and is a motivation to extract a set of biologically meaningful EFMs according to a biological aspect such as motifs [23] and thermodynamics [24].

The main focus of this paper is to introduce a data model based on the AND/OR graph and to propose an approach to find flux-balanced pathways according to pathway topology and reaction stoichiometries. It is shown that the computed flux-balanced pathways are a biologically relevant subset of EFMs which include external input and output metabolites. Besides, based on the introduced graph data model, an upper-bound for the complexity of exploring EFMs containing external metabolites is calculated. Using the topology of the network gives us the opportunity to make decisions according to the metabolite/reaction aspects to preserve or eliminate a particular reaction or metabolite in a certain path. Introducing a model to consider both *topology* and *stoichiometry* of a metabolic network for network analysis may lead to better biological decisions and output categorization.

The rest of the text is organized as follows. Some required concepts and preliminaries are provided in Section 2. Before explaining the proposed approach to find the elementary flux modes in a given metabolic network (Section 3.2), we first describe the modified AND/OR graph model and its properties in Section 3.1. Finally, Section 4 is devoted to results and discussions and Section 5 concludes the paper.

2 Preliminaries

In this section, some basic concepts along with the formal definition of elementary flux modes are provided.

2.1 Metabolic Networks

Metabolic networks model the metabolism of living cells in terms of a set of biochemical reactions. Definitions in this section are derived from [25]. The biochemical reactions can be irreversible, i.e., the reaction can be active only in one direction, or reversible, i.e., the reaction can be active in both directions. The contributing metabolites in a reaction can be either substrates or products. Substrates are consumed and products are produced during the operation of a reaction. The topology of a metabolic network is characterized by its $m \times n$ stoichiometric matrix, S , where m and n correspond to the number of metabolites and reactions, respectively. In this paper, external metabolites are not included in S . The value S_{ij} represents the stoichiometric coefficient of the metabolite i in the reaction j . S_{ij} is positive/negative if the metabolite i is produced/consumed. If this coefficient is zero it means that the metabolite i does not contribute to the reaction j . The network is considered in the steady-state if for each internal metabolite, the rates of consumption and production are equal. The reactions connected to the external metabolites are called *Boundary* reactions.

2.2 Elementary Flux Mode Definition

A flux vector $\mathbf{v} \neq \mathbf{0}$ and $\mathbf{v} \in R^n$ is considered an EFM if it meets the following conditions:

- $v_i \geq 0$ for all $i \in$ irreversible reactions (thermodynamic constraint).
- $\mathbf{S} \cdot \mathbf{v} = \mathbf{0}$ (steady-state condition).
- There is no $\mathbf{v}' \in R^n$ with $\text{supp}(\mathbf{v}') \subset \text{supp}(\mathbf{v})$, where support of a mode is defined as $\text{supp}(\mathbf{v}) = \{i | v_i \neq 0\}$, (minimality limitation).

3 Methods

In this section, the proposed graph-based elementary flux mode analysis is discussed in detail. To do so, the data structure used in our approach is defined, and then, the proposed approach is introduced accordingly.

3.1 Data Model

The proposed data model is based on the conventional AND/OR graph in computer science. However, we provide a different definition with additional features to make the model appropriate for our proposed algorithm. In our model, the coefficients of the metabolites in reactions are embedded in the graph structure as attributes of each node (explained in Definition 3, see below). A *pathway* is defined as a set of reactions with their associated fluxes. The flux of each reaction in a certain path is also embedded in the graph.

Definition 1. The Modified Graph for representing a metabolic network, denoted as **MG**, is defined as a set of *Nodes*, \mathcal{N}_i , i.e., $\mathbf{MG} = \{\mathcal{N}_i | 0 \leq i \leq M - 1\}$, where M is the number of internal metabolites in the metabolic network.

Definition 2. Each node \mathcal{N}_i in **MG** is a 3-tuple $\mathcal{N} = (i, I, O)$ where

- i is the tag of a metabolite,
- I is an array of input reactions that produce the metabolite and
- O is an array of output reactions that consume the metabolite.

Definition 3. Each I/O in Definition 2 contains the following data:

- The reaction j , $0 \leq j \leq r - 1$, where r is the number of reactions consuming/producing the metabolite i ,
- I_{M_j}/O_{M_j} , an array of the metabolites consumed/produced by reaction j . In other words, $I_{M_j}/O_{M_j} = \{m_{kj} | 0 \leq k \leq m - 1\}$, where m is the number of consumed/produced metabolites by the reaction j ,
- The direction of the reaction for reversible reactions,
- $I_{c_{ij}}/O_{c_{ij}}$, the coefficient of the reaction j for the produced/consumed metabolite i in the stoichiometric matrix S and
- $I_{f_{ijp}}/O_{f_{ijp}}$, the flux of the input/output reaction j for a certain path p .

A general configuration of a node \mathcal{N}_i in **MG** is shown in Figure 1. The incoming arcs (i.e., I in Definition 2) in each \mathcal{N}_i produce the metabolite i and the outgoing arcs (i.e., O in Definition 2) consume it. Therefore, consumed metabolites, m_i , and produced metabolites, m'_i , contributing to reaction j as shown in Equation 1, are as arcs between \mathcal{N}_i nodes each associated to one metabolite. I arcs and O arcs in \mathcal{N}_i nodes are related to each other by reaction tags in Definition 3.

$$r_j : m_1 + m_2 + \dots + m_i + \dots \rightleftharpoons m'_1 + m'_2 + \dots + m'_i + \dots \quad (1)$$

3.2 GB-EFM: Graph-Based EFM Analysis Algorithm

The proposed algorithm for finding EFMs of a metabolic network is based on the following facts:

1. Starting from boundary metabolites, each incoming flux in a path produce some metabolite(s), and each produced metabolite should be consumed and the flux of the contributed reactions can be obtained according to the stoichiometric coefficients in order to make sure that the constructed path is in steady-state.

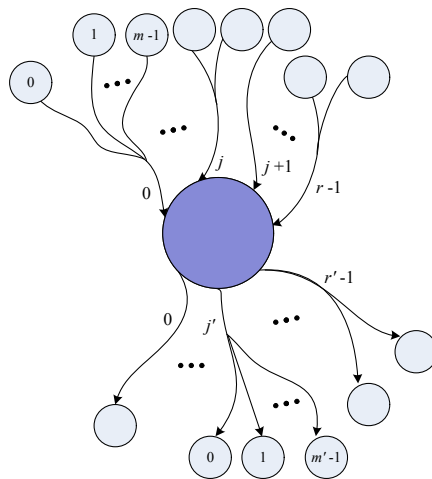


Figure 1: A general configuration of a node \mathcal{N}_i in a modified AND/OR graph as defined in Definitions 1, 2 and 3.

2. To guarantee the elementarity of the produced pathways, two rules are followed: (1) Only one output from the node \mathcal{N}_i should be considered when constructing a path and (2) multi-path condition should be checked in a node which has more than one input and one output. The multi-path condition occurs in a node when independent paths share part of a path while traversing the graph.
3. When one metabolite is consumed by a reaction, the presence of its AND-related metabolites are required. For example, when metabolites A and B are AND-related, if A is consumed during the reaction $A+B \rightarrow C+D$, then B should also be present which means A and B should be consumed simultaneously. Therefore, the algorithm is designed to have a forward/backward flow for the produced/consumed metabolites. In this example, the metabolites A and B are consumed while C and D are produced.

The proposed graph-based EFM analysis algorithm is composed of the following five steps:

Step 0. Construction of MG. The graph **MG** is constructed based on the stoichiometric matrix as stated in Algorithm 1.

Step 1. Finding independent paths with dependent nodes. In this step, starting from the metabolites connected to the boundary reactions, all possible paths are constructed by traversing the internal metabolites. Proposition 1 shows that the paths are constructed with minimum number of possible reactions using Definition 4 and Definition 5. The semi-minimality occurs since different paths are constructed from different outputs and each node is visited

ALGORITHM 1: Construction of Modified AND/OR Graph.

Input: S : The stoichiometric matrix of the internal metabolites in a metabolic network and an array V representing reversible reactions
Output: The equivalent **MG** of S

```
for all metabolites  $i$ ,  $0 \leq i \leq M - 1$ , the rows of  $S$  do
  for all reactions  $j$ ,  $0 \leq j \leq R - 1$ , the columns of  $S$  do
    Create node  $\mathcal{N}_i$ .
    for all irreversible reactions do
      if  $S_{ij} > 0$  then
        | add an input to  $\mathcal{N}_i$  with the reaction tag  $j$ .
      else if  $S_{ij} < 0$  then
        | add an output to  $\mathcal{N}_i$  with the reaction tag  $j$ .
      end
    end
    for all reversible reactions  $v_i \in V$  do
      | Add both an input and an output to  $\mathcal{N}_i$  with the reaction tag  $j$ .
    end
  end
end
```

multiple times only if required.

Definition 4. Forward/Backward Flow. The *forward* flow is applied for the metabolites of the output reaction and the *backward* flow is applied for the and-related metabolites of the input reaction.

Definition 5. Primary and Secondary Reactions. On each path, each node has a *primary* input and a *primary* output which directs the flow of the path and tags the metabolite as *visited*. When an edge (i.e., a reaction) enters an already visited node in that path in a forward/backward flow as an input/output, that reaction is tagged as a *secondary* input/output of that node.

Proposition 1. *Starting from metabolites connected to the boundary reactions, r paths are constructed from each node where r is the number of output/input reactions of that node in the forward/backward flow. The paths are semi-minimal based on the following statements:*

- *each output/input reaction in forward/backward flow is traversed once and labeled as primary, and*
- *when an edge enters an already visited node, the path is considered as closed in the node and the edge is labeled as secondary.*

In the forward flow of a node, for each output reaction, an independent path is constructed. Each path is traversed independently of the other paths. In the backward flow of a node, for each input reaction an independent path is constructed. In Figure 2 the forward/backward flow of each node \mathcal{N}_i is shown in an illustrative example. The information of each path is saved independently.

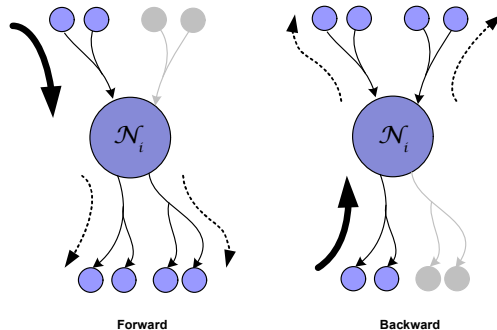


Figure 2: Constructing possible paths from outputs/inputs in forward/backward flow in Step 1. In the forward/backward flow, from the primary input/output illustrated by the thick arrow, two independent paths are constructed along each output/input reaction illustrated by dashed arrows.

Step 2. Eliminate multi-way paths. By using Proposition 2 and eliminating multi-way paths (defined in Definition 7) which include nodes with multi-path condition (defined in Definition 6), the set of semi-minimal paths is reduced to the set of minimal paths. An example of this is shown in Figure 3.

Definition 6. Multi-Path Condition. This is a condition that occurs in a node with more than one input and more than one output and two independent paths from different origins share this node.

Definition 7. Multi-Way Paths. They are semi-minimal paths that contain reactions from two independent paths are referred to as multi-way paths.

Proposition 2. *Since different paths are constructed from different outputs of a node and there are both forward and backward flows, depending on the order of traversing the nodes, there exist multi-way paths containing reactions from two independent paths. The path is minimal if it is not a multi-way path. In other words, it has no node with multi-path condition applied to it.*

To solve the multi-path condition, for each produced path in Step 1, we traverse the subgraph of that path and tag each reaction with its associated metabolites. Since the structure of the path is known, one can label the outgoing/ingoing reactions in forward/backward flow of a path.

Then, in a path, for each node with more than one input and more than one output reaction, if one can find two routes crossing the node and the origins of the routes are different, the path should be discarded. This is checked according to the labels, i.e., $[node.pathNum]$, which has been assigned to each reaction. $pathNum$ is a variable which counts the number of paths generated by a node. In other words, this path can be functional by either of the found routes in the node.

Step 3. Adjusting reaction fluxes. In this step, every node i in a path p is examined and the fluxes of the primary input and the primary output of that node are calculated using the following equations:

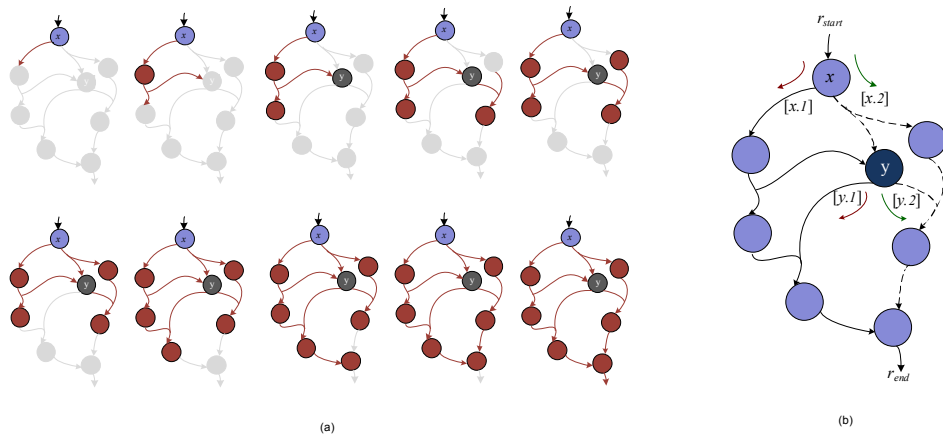


Figure 3: An example of the multi-path condition and a multi-way path. (a) This figure shows how the order of traversing the nodes may cause a multi-path condition in node y . (b) In general, starting from node x , two paths are constructed and their reactions are labeled as $x.1$ and $x.2$. After arriving at node y from each of the paths, two other paths are constructed and their reactions are labeled as $y.1$ and $y.2$. Altogether four different paths are obtained, including the reaction pairs $([x.1],[y.1])$, $([x.1],[y.2])$, $([x.2],[y.1])$ and $([x.2],[y.2])$. As can be seen in the graph, the paths with reactions labeled with 1 are converged to each other and the paths which their reactions labeled with 2 are converged to each other as well. Therefore, two of the four paths passing $([x.1],[y.2])$, $([x.2],[y.1])$ in fact belong to the same metabolic pathway but the two other paths, path 1 shown by solid lines and path 2 shown by dashed lines, are independent. The starting and ending reactions belong to both paths. Step 2 identifies this by labeling the outputs and finding the multi-path nodes to eliminate multi-way paths which are a combination of independent paths.

$$\begin{aligned}
O_{f_{ijp}} &= \frac{I_{f_{ijp}} I_{c_{ij}}}{O_{c_{ij}}}, & \text{Forward} \\
I_{f_{ijp}} &= \frac{O_{f_{ijp}} O_{c_{ij}}}{I_{c_{ij}}}, & \text{Backward}
\end{aligned} \tag{2}$$

In Equation 2, j refers to the primary input/output reaction. Proposition 3 explains how all reactions in a path p get a flux.

Proposition 3. *Starting from a boundary reaction with a certain flux, all nodes in a path get a flux using Equation 2, as each node has a primary input and a primary output to be used in this equation. Equation 2 states that the amount of the incoming and outgoing fluxes in \mathcal{N}_i should be equal.*

Step 4. Balancing flux. After Step 3, the consumption/production balance of the nodes with secondary reactions may be disturbed. To fix this, we go through these nodes and find a way forward/backward to output/input reactions to recalculate the extra production/consumption of the unbalanced nodes. Using Proposition 4, the paths with no balance rate are discarded. If $\sum I_{c_{ij}} I_{f_{ijp}} - \sum O_{c_{ij}} O_{f_{ijp}} = 0$, the balance condition is accepted for this node. Otherwise, on this path, the extra flux is added to the next reactions on the path, towards the boundary reactions using Equation 3. In this case, the difference between the input rates and the output rates is equal to $I_{c_{ie}} I_{f_{iep}}$ or $O_{c_{ie}} O_{f_{iep}}$ where e implies the reaction which causes the extra production/consumption of the \mathcal{N}_i . The old/new value of the flux is shown by o/n labels in Equation 3 and j refers to the primary input/output reaction.

$$\begin{aligned}
I_{f(n)_{ijp}} &= \frac{I_{f(o)_{ijp}} I_{c_{ij}} - I_{f_{iep}} I_{c_{ie}}}{I_{c_{ij}}}, \\
I_{f(n)_{ijp}} &= \frac{I_{f(o)_{ijp}} I_{c_{ij}} + O_{f_{iep}} I_{c_{ie}}}{I_{c_{ij}}}, \\
O_{f(n)_{ijp}} &= \frac{O_{f(o)_{ijp}} O_{c_{ij}} + O_{f_{iep}} I_{c_{ie}}}{O_{c_{ij}}}, \\
O_{f(n)_{ijp}} &= \frac{O_{f(o)_{ijp}} O_{c_{ij}} - I_{f_{iep}} I_{c_{ie}}}{O_{c_{ij}}}
\end{aligned} \tag{3}$$

Proposition 4. *Flux dependencies can prevent the rate of the production/consumption of an internal metabolite in a given topology of a path from being zero. Equation 3 is used to calculate an update for the consumption/production rate of the node. In the case of*

- $I_{f(n)_{ijp}} = 0$ or $O_{f(n)_{ijp}} = 0$ or
- a repetitive loop over a node (i.e., getting back to a node from the same reaction multiple times), while trying to find a way out in a subgraph of a given path,

the path is discarded.

ALGORITHM 2: Graph-Based EFM Analysis Algorithm.

Input: The **MG** graph of a metabolic network derived from its given stoichiometric matrix.

Output: The set of elementary flux modes of the given network from input metabolites to output metabolites.

Store all boundary metabolites of the network with their associated reactions in the queue \mathcal{Q} .

```

for each entry  $\mathcal{Q}_i$  in the queue,  $0 \leq i \leq \text{size}(\mathcal{Q}) - 1$  do
    for each  $O_j/I_j$ , in forward/backward flow do
        Start a new path for each reaction  $j$  as a primary output/input.
        if  $\mathcal{N}_i \notin \mathcal{P}$  then
            Add  $\mathcal{N}_i$  to  $\mathcal{P}$  accordingly for each path.
        else
            Add  $j$  to  $\mathcal{N}_i$  as a secondary output/input.
        end
        Push back  $O_{M_i}$  to  $\mathcal{Q}$  to be considered in the forward path of the algorithm.
        Push back  $I_{M_i}$ , except for the parent node, to  $\mathcal{Q}$  to be considered in the backward path
        of the algorithm.
        Repeat until  $\mathcal{Q}=\emptyset$ .
    end
end

for all  $\mathcal{P}_k$ ,  $0 \leq k \leq P - 1$  do
    for all  $\mathcal{N}_i$  in  $\mathcal{P}_k$  do
        if  $j \neq 1$  then
            for all  $O_j/I_j$ ,  $0 \leq j \leq r - 1$  in forward/backward flow do
                Add label  $\mathcal{N}_{ij}$  to  $O_j/I_j$ .
            end
        else
            Pass the label from input/output to output/input in forward/backward flow.
        end
    end
    for all  $\mathcal{N}_i$  in  $\mathcal{P}_k$  with more than two inputs and two outputs do
        if  $\mathcal{N}_x \in \bigcup_M \mathcal{N}_m$  where  $M$  contains all visited metabolites associated with the reaction  $j$ 
        then
            if a cross match for  $\mathcal{N}_{xy}$  and  $\mathcal{N}_{xz}$  in the inputs is found as well as a match in the
            outputs then
                Discard  $\mathcal{P}_k$ .
            end
        end
    end
end

for all remaining  $\mathcal{P}_k$ ,  $0 \leq k \leq P - 1$  do
    for all  $\mathcal{N}_i$  in  $\mathcal{P}_k$  do
        Apply Eq. 1.
    end
end

for all remaining  $\mathcal{P}_k$ ,  $0 \leq k \leq P - 1$  do
    for all  $\mathcal{N}_i$  in  $\mathcal{P}_k$  with Secondary inputs/outputs do
        Calculate  $\sum_j I_{f_{iek}} I_{c_{ie}} = \sum_j I_{f_{ijk}} I_{c_{ij}} - \sum_j O_{f_{ijk}} O_{c_{ij}}$ .
        if  $\sum_j I_{f_{iek}} I_{c_{ie}} \neq 0$  then
            Use Eq. 2 to pass the flux to the primary input/output reaction of  $\mathcal{N}_i$ .
            Repeat until reaching the boundary reactions.
        end
    end
end

```

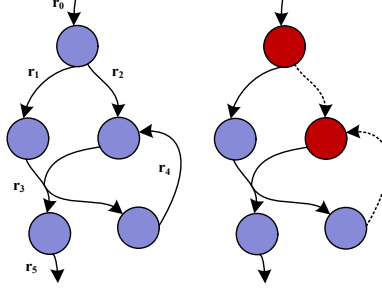


Figure 4: An illustration of Steps 3 and 4 on a sample graph. The pairs (r_0, r_1) , (r_1, r_3) , (r_3, r_5) , (r_3, r_2) , (r_3, r_4) act as primary reactions (input, output) for the five nodes in Step 3. The two nodes with dashed secondary reactions update production/consumption of metabolites for r_2/r_4 in Step 4.

The proposed GB-EFM analysis algorithm is presented in Algorithm 2 and is summarized in Theorem 1.

Theorem 1. *Starting from boundary nodes in a metabolic network, GB-EFM analysis algorithm produces all minimal flux-balanced pathways connecting input and output metabolites with the following properties:*

- $R_k \not\subset R_l$, for all \mathcal{P}_k and \mathcal{P}_l , $k, l \in [0, P-1]$ $k \neq l$.
- $\sum_j I_{f_{ijk}} I_{c_{ij}} = \sum_j O_{f_{ijk}} O_{c_{ij}}$, for all nodes \mathcal{N}_i , $0 \leq i \leq M-1$.

where \mathcal{P}_k refers to a path $0 \leq k \leq P-1$ and R_k is the set of contributing reactions in the path. The above properties indicate that the produced paths are elementary flux modes.

Proof. Step 0 provides the required interface from the stoichiometric matrix. By using Propositions 1 and 2, semi-minimal and then minimal pathways are produced, respectively. According to Propositions 3 and 4, steady-state condition is checked for the given paths and fluxes are calculated. Non-qualified paths are discarded in Steps 2 and 4 based on Propositions 2 and 4. The remaining paths have minimum functional reactions with balanced fluxes as elementary flux modes. Therefore, the produced paths are a subset of EFMs connecting input and output external metabolites. \square

A complete framework of Algorithm 2 is illustrated in Figure 5.

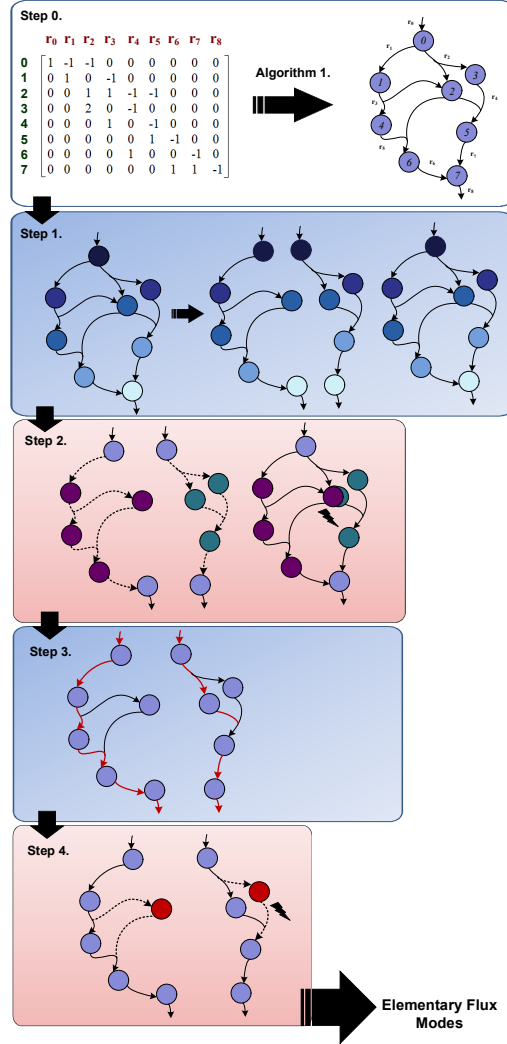


Figure 5: An example illustrating Steps 0 to 4 on a simple network. In Step 1, three different paths are constructed from a given **MG** produced in Step 0. The paths are divided as we go further in the graph, as shown by dark-colored nodes changing to light-colored nodes. In Step 2, for the first two paths from left, there is no node with more than one input and one output. Therefore, multi-path condition is not applied. However, the third path is constructed from the first two paths. The internal nodes of each of these two paths are shown by different colors. Multi-path condition is checked in their common node shown by overlapped circles. As a result, this path is discarded. In Step 3, for each path, all nodes are traversed and primary inputs and primary outputs, illustrated by red solid lines, get a flux. In Step 4, nodes with secondary input(s)/output(s) are traversed again to pass the extra flux to the boundary reactions. Secondary input(s)/output(s) are illustrated by dotted lines and their associated nodes are illustrated by red circles. According to the coefficients, as shown in the given stoichiometric matrix embedded in the graph, the second path cannot get a stable flux value for all reactions and is discarded. The final output after Step 4 is one EFM.

4 Results and Discussion

4.1 Computational Complexity

To calculate the computational complexity of Algorithm 2, the complexity of each step is analyzed below:

- In Step 1, the maximum number of created paths is calculated as $P_{\max} = M_B \times (r_{\max})^{M \setminus B}$, in which M_B is the number of boundary metabolites, $M \setminus B$ is the number of all other internal metabolites except the boundary ones and r_{\max} is the maximum number of input/output reactions for all nodes.
- In Step 2, for each path \mathcal{P}_i , there are two traversals with the complexity of (1) $M_{\mathcal{P}_i}$, the number of all nodes in a path, and (2) $M_{t\mathcal{P}_i}$, the number of all tagged nodes with more than one input and one output in that path. Therefore, the order of maximum traversals is $P \times [M_{x\mathcal{P}_i} \times M_{\mathcal{P}_i}]$ or $P \times M^2$ where $M_{\mathcal{P}_i}$ and $M_{t\mathcal{P}_i} \leq M$.
- In Step 3, all nodes in \mathcal{P}_i should be traversed which requires $P' \times M_{\mathcal{P}_i}$ iterations. $P' \leq P$ because some paths may be discarded in Step 2. Therefore, the worst-case computational complexity for this step would be $P \times M$.
- In Step 4, for each \mathcal{P}_i and for all nodes with secondary reactions, (i.e., $M_{v\mathcal{P}_i}$ nodes), at least half of the nodes should be traversed to lead the extra production/consumption of metabolites to input/output. Therefore, $P' \times [M_{v\mathcal{P}_i} \times M_{\mathcal{P}_i/2}]$ is the computational complexity of this step. Since $P' \leq P$ and $M_{v\mathcal{P}_i}, M_{\mathcal{P}_i/2} \leq M$, $P \times [M^2/2]$ is the computational complexity in the worst-case.

Considering all items above, $P \times O(M^2)$, or as a result $r_{\max}^M \times O(M^2)$ is the computational complexity of the algorithm if we assume $M_B \ll M$. The computational complexity of the approach can be considered as an upper-bound for the number of EFMs including external reactions.

4.2 EFM Topology Description

Three different topologies for an EFM can be observed in metabolic networks:

- Internal paths with no boundary reaction(s),
- paths with only input or only output reaction(s) consisting of an internal loop,
- paths from input reaction(s) to output reaction(s).

These topologies are illustrated in Figure 6. The main focus of Algorithm 2 is the third topology. However, in order to expand this focus to all EFMs,

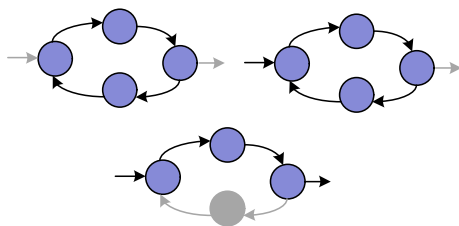


Figure 6: Categorization of EFM topologies

the algorithm can be easily modified to keep paths without the contribution of output reactions (in Case 1) rather than discarding them, or starting from internal reactions (in Case 2).

However, from the biological point of view, internal paths with no boundary reactions, (i.e., loops) are called “futile cycles” and are not biotechnologically relevant [26], and the paths with boundary inputs and no output implicate inconsistency in the network [27]. Therefore, a meaningful subset of EFMs are targeted here to keep more relevant EFMs in the output.

4.3 Application to Metabolic Network Models

To demonstrate the functionality of the model, the GB-EFM method has been implemented in C++ and tested on an Intel Core-i5 CPU with 4 GB RAM. The source code and results are provided on <https://github.com/marabzadeh/GB-EFM>. A small metabolic network comprising tricarboxylic acid cycle, glyoxylate shunt and adjacent reactions of amino acid synthesis in *E. coli* [28] is considered to describe the steps of the algorithm and proof the concept. The network has one boundary input reaction and three boundary output reactions among all 24 reactions and 18 internal metabolites as shown in Table 1. Reactions R_7 , R_{12} and R_{13} are boundary output reactions. The elementary flux modes were obtained using EFMtool [16] as well as the proposed method. EFMtool gives 12 elementary flux modes in which 3 are inside loops, 4 contain only an input boundary reaction without any output reaction and the other 5 are pathways from the input to all outputs.

The number of paths produced by the proposed algorithm after Step 1 is 40. After Step 2, this is reduced to 30 and after applying Steps 3-4, 5 pathways (final EFMs) remain. These pathways are the same as the five “acyclic” EFMs obtained by EFMtool. The resulting EFMs are illustrated in Figure 7.

Besides, a set of moderate-sized metabolic networks are considered as further test cases. The first network is the pentose phosphate pathway in *trypanosoma brucei* [30], the second one is the simple network of human red blood cell metabolism [31] and the third one is an *E. coli* core model obtained from [32]. GB-EFM found all EFMs from *Glucose* as reported in Table 2. Computational time for all the networks was less than one second. The currency metabolites such as ADP, ATP, AMP, CO_2 , H_2O , O_2 , H_2 , NH_4 , Pi, NAD, NADH, NADP

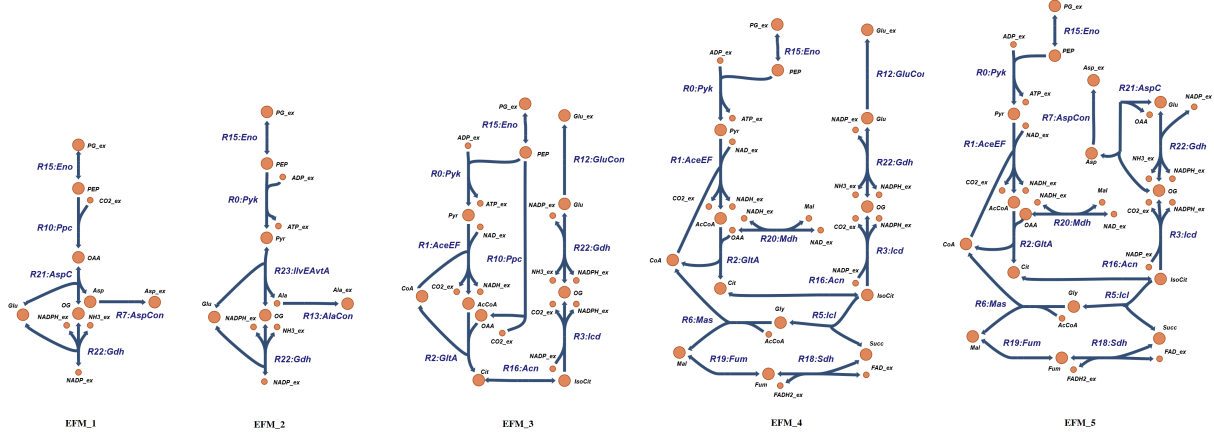


Figure 7: The resulting EFMs are depicted using Escher [29].

Table 1: The list of reactions for the sample network of *E. coli* model and the resulting EFMs obtained by GB-EFM. External output metabolites and cofactors are marked by asterisks. The external metabolite consumed by the input boundary reaction R_{15} is marked by two asterisks.

Reactions					EFMs				
#	Name	Cons	Dir	Prod	1	2	3	4	5
R0	Pyk	PEP + ADP*	\Rightarrow	Pyr + ATP*	0	1	1	3	2
R1	AceEF	Pyr + NAD* + CoA	\Rightarrow	AcCoA + CO ₂ * + NADH*	0	0	1	3	2
R2	GltA	OAA + AcCoA	\Rightarrow	Cit + CoA	0	0	1	2	1
R3	Icd	IsoCit + NADP*	\Rightarrow	OG + CO ₂ * + NADPH*	0	0	1	1	0
R4	SucAB	OG + NAD* + CoA	\Rightarrow	SucCoA + CO ₂ * + NADH*	0	0	0	0	0
R5	Icl	IsoCit	\Rightarrow	Succ + Gly	0	0	0	1	1
R6	Mas	Gly + AcCoA	\Rightarrow	Mal + CoA	0	0	0	1	1
R7	AspCon	Asp	\Rightarrow	Asp.ex*	1	0	0	0	1
R8	AspA	Asp	\Rightarrow	Fum + NH ₃ *	0	0	0	0	0
R9	Pck	OAA + ATP*	\Rightarrow	PEP + ADP* + CO ₂ *	0	0	0	0	0
R10	Ppc	PEP + CO ₂ *	\Rightarrow	OAA	1	0	1	0	0
R11	Pps	Pyr + ATP*	\Rightarrow	PEP + AMP*	0	0	0	0	0
R12	GluCon	Glu	\Rightarrow	Glu.ex*	0	0	1	1	0
R13	AlaCon	Ala	\Rightarrow	Ala.ex*	0	1	0	0	0
R14	SucCoACon	SucCoA	\Rightarrow	Suc.ex* + CoA	0	0	0	0	0
R15	Eno	PG**	\Leftrightarrow	PEP	1	1	2	3	2
R16	Acn	Cit	\Leftrightarrow	IsoCit	0	0	1	2	1
R17	SucCD	SucCoA + ADP*	\Leftrightarrow	Succ + ATP* + CoA	0	0	0	0	0
R18	Sdh	Succ + FAD*	\Leftrightarrow	Fum + FADH ₂ *	0	0	0	1	1
R19	Fum	Fum	\Leftrightarrow	Mal	0	0	0	1	1
R20	Mdh	Mal + NAD*	\Leftrightarrow	OAA + NADH*	0	0	0	2	2
R21	AspC	OAA + Glu	\Leftrightarrow	Asp + OG	1	0	0	0	1
R22	Gdh	OG + NH ₃ * + NADPH*	\Leftrightarrow	Glu + NADP*	1	1	1	1	1
R23	IlvEAvtA	Pyr + Glu	\Leftrightarrow	Ala + OG	0	1	0	0	0

Table 2: Number of EFMs starting from *Glucose* in some networks.

Network	Size ($m * n$)	#EFMs
Glycolysis and Pentose Phosphate pathway in <i>T. brucei</i>	26*35	4
Human red blood cell	20*50	20
<i>E. coli</i> core (Escherichia coli str. K-12 substr. MG1655)	53*94	47

and NADPH were removed from the input stoichiometry matrix of all samples.

4.4 Discussion

According to the graph data model, the order of traversing the nodes and the relations between reactions and metabolites can be maintained for each path. Besides, decisions to select or eliminate a particular reaction or metabolite can be made during the algorithm. Since in this approach the traversal is on the graph, many paths are produced which may be discarded afterwards.

The proposed data model gives the opportunity of different levels of parallelism to trace pathways. Therefore, the complexity of implementation can be reduced by optimizing the method to only focus on paths that are desirable in EFM analysis and by exploiting advantages of the possible parallelism in the method.

The proposed framework can be considered as an extension to the well-known flow network problem for hypergraphs. The edges are directed and each has been associated with a weight. However, the weight of an output arc of a node is different from the weight when the arc enters another node as an input. Besides, the hypergraph structure complicates pathways topology [33].

The first introduced double-description method for finding EFMs explores the whole set of pathways to find paths that are in the steady-state solution space and then selects EFMs by comparing the reaction subset of pathways. In the improved double-description versions of the method which use null-space of the stoichiometry matrix as an input, different combinations of pathways in the null-space are calculated and then the reaction subset of pathways are compared. The main difference of the GB-EFM method with double-description-based methods is that GB-EFM first calculates the shortest pathways in the solution space and then checks to see if these pathways can be in the steady-state condition by using some rules on the topology of the pathway.

5 Conclusion and Future Work

In order to calculate the elementary flux modes of a given metabolic network, an algorithm was proposed based on a modified AND/OR graph. It was shown in the paper that the steps in the algorithm leads to exact pathways according to EFM definition. The worst-case computational complexity of the algorithm was calculated as $O(r_{\max}^M \times M^2)$ which is a newly reported upper-bound for

the number of EFMs with external metabolites. Additionally, a set of test cases was provided to prove the concept of the algorithm.

The main focus of this paper was to introduce a model to find minimal flux-balanced pathways in metabolic networks. The model can also be applicable to several constraint-based pathway analysis approaches. Using the potential parallelism in the method to speed-up the algorithm on a hardware structure and proposing different categorization of EFMs based on reactions/metabolites characteristics, as is possible according to the graph data model, are considered as our future research.

Acknowledgements

Authors would like to thank Dr. Nathan Lewis (UCSD) for helpful discussions on the concept.

References

- [1] C. S. Henry, M. DeJongh, A. A. Best, P. M. Frybarger, B. Linsay, R. L. Stevens, High-throughput generation, optimization and analysis of genome-scale metabolic models, *Nature biotechnology* 28 (9) (2010) 977–982.
- [2] N. D. Price, J. L. Reed, B. Ø. Palsson, Genome-scale models of microbial cells: evaluating the consequences of constraints, *Nature Reviews Microbiology* 2 (11) (2004) 886–897.
- [3] J. D. Orth, I. Thiele, B. Ø. Palsson, What is flux balance analysis?, *Nature biotechnology* 28 (3) (2010) 245–248.
- [4] E. Klipp, R. Herwig, A. Kowald, C. Wierling, H. Lehrach, *Systems biology in practice: concepts, implementation and application*, John Wiley & Sons, 2008.
- [5] S. Schuster, C. Hilgetag, On elementary flux modes in biochemical reaction systems at steady state, *Journal of Biological Systems* 2 (02) (1994) 165–182.
- [6] S. Schuster, D. A. Fell, T. Dandekar, A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks, *Nature biotechnology* 18 (3) (2000) 326–332.
- [7] S. Schuster, S. Klamt, W. Weckwerth, F. Moldenhauer, T. Pfeiffer, Use of network analysis of metabolic systems in bioengineering, *Bioprocess and Biosystems Engineering* 24 (6) (2002) 363–372.
- [8] D. Radhakrishnan, M. Rajvanshi, K. Venkatesh, Phenotypic characterization of *Corynebacterium glutamicum* using elementary modes towards synthesis of amino acids, *Systems and synthetic biology* 4 (4) (2010) 281–291.
- [9] K. Parvatham, L. Veerakumari, Drug target prediction using elementary mode analysis in *Ascaris lumbricoides* energy metabolism, *Biotechnology and bioprocess engineering* 18 (3) (2013) 491–500.
- [10] D. Machado, M. J. Herrgård, Co-evolution of strain design methods based on flux balance and elementary mode analysis, *Metabolic Engineering Communications* 2 (2015) 85–92.

- [11] J. A. Papin, N. D. Price, S. J. Wiback, D. A. Fell, B. O. Palsson, Metabolic pathways in the post-genome era, *Trends in biochemical sciences* 28 (5) (2003) 250–258.
- [12] C. Wagner, Nullspace approach to determine the elementary modes of chemical reaction systems, *The Journal of Physical Chemistry B* 108 (7) (2004) 2425–2431.
- [13] R. Urbanczik, C. Wagner, An improved algorithm for stoichiometric network analysis: theory and applications, *Bioinformatics* 21 (7) (2005) 1203–1210.
- [14] L. E. Quek, L. K. Nielsen, A depth-first search algorithm to compute elementary flux modes by linear programming, *BMC systems biology* 8 (1) (2014) 1.
- [15] J. Gagneur, S. Klamt, Computation of elementary modes: a unifying framework and the new binary approach, *BMC bioinformatics* 5 (1) (2004) 1.
- [16] M. Terzer, J. Stelling, Large-scale computation of elementary flux modes with bit pattern trees, *Bioinformatics* 24 (19) (2008) 2229–2235.
- [17] A. Von Kamp, S. Schuster, Metatool 5.0: fast and flexible elementary modes analysis, *Bioinformatics* 22 (15) (2006) 1930–1931.
- [18] L. F. De Figueiredo, A. Podhorski, A. Rubio, C. Kaleta, J. E. Beasley, S. Schuster, F. J. Planes, Computing the shortest elementary flux modes in genome-scale metabolic networks, *Bioinformatics* 25 (23) (2009) 3158–3165.
- [19] L. David, A. Bockmayr, Computing elementary flux modes involving a set of target reactions, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 11 (6) (2014) 1099–1107.
- [20] J. F. H. Céspedes, F. D. A. G. Asensio, J. M. G. Carrasco, A new approach to obtain efms using graph methods based on the shortest path between end nodes, in: *International Conference on Bioinformatics and Biomedical Engineering*, Springer, 2015, pp. 641–649.
- [21] E. Ullah, S. Aeron, S. Hassoun, gEFM: an algorithm for computing elementary flux modes using graph traversal, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 13 (1) (2016) 122–134.
- [22] S. Klamt, J. Stelling, Combinatorial complexity of pathway analysis in metabolic networks, *Molecular biology reports* 29 (1-2) (2002) 233–236.
- [23] S. Pérès, F. Vallée, M. Beurton-Aimar, J.-P. Mazat, Acom: a classification method for elementary flux modes based on motif finding, *BioSystems* 103 (3) (2011) 410–419.
- [24] M. P. Gerstl, D. E. Ruckerbauer, D. Mattanovich, C. Jungreuthmayer, J. Zanghellini, Metabolomics integrated elementary flux mode analysis in large metabolic networks, *Scientific reports* 5 (2015) 8930.
- [25] J. Zanghellini, D. E. Ruckerbauer, M. Hanscho, C. Jungreuthmayer, Elementary flux modes in a nutshell: properties, calculation and applications, *Biotechnology journal* 8 (9) (2013) 1009–1016.
- [26] A. P. Burgard, S. Vaidyaraman, C. D. Maranas, Minimal reaction sets for *Escherichia coli* metabolism under different growth requirements and uptake environments, *Biotechnology progress* 17 (5) (2001) 791–797.
- [27] A. Gevorgyan, M. G. Poolman, D. A. Fell, Detection of stoichiometric inconsistencies in biomolecular models, *Bioinformatics* 24 (19) (2008) 2245–2251.

- [28] S. Schuster, T. Dandekar, D. A. Fell, Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering, *Trends in biotechnology* 17 (2) (1999) 53–60.
- [29] Z. A. King, A. Dräger, A. Ebrahim, N. Sonnenschein, N. E. Lewis, B. O. Palsson, Escher: a web application for building, sharing, and embedding data-rich visualizations of biological pathways, *PLoS Comput Biol* 11 (8) (2015) e1004321.
- [30] E. J. Kerkhoven, F. Achcar, V. P. Alibu, R. J. Burchmore, I. H. Gilbert, M. Trybilo, N. N. Driessen, D. Gilbert, R. Breitling, B. M. Bakker, et al., Handling uncertainty in dynamic models: the pentose phosphate pathway in trypanosoma brucei, *PLoS Comput Biol* 9 (12) (2013) e1003371.
- [31] S. J. Wiback, B. O. Palsson, Extreme pathway analysis of human red blood cell metabolism, *Biophysical Journal* 83 (2) (2002) 808–818.
- [32] Z. A. King, J. Lu, A. Dräger, P. Miller, S. Federowicz, J. A. Lerman, A. Ebrahim, B. O. Palsson, N. E. Lewis, Bigg models: A platform for integrating, standardizing and sharing genome-scale models, *Nucleic acids research* 44 (D1) (2016) D515–D522.
- [33] S. A. Marashi, M. Tefagh, A mathematical approach to emergent properties of metabolic networks: Partial coupling relations, hyperarcs and flux ratios, *Journal of theoretical biology* 355 (2014) 185–193.